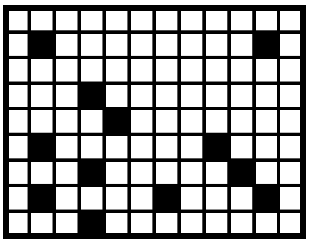
7. ΣΥΓΚΡΙΣΗ ΚΑΙ ΣΥΝΔΙΑΣΜΟΣ ΤΩΝ ΔΙΑΦΟΡΩΝ ΜΕΘΟΔΩΝ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

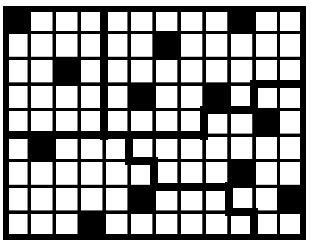
7.1. ΣΥΓΚΡΙΣΗ ΤΩΝ ΒΑΣΙΚΩΝ ΣΤΡΑΤΗΓΙΚΩΝ

Στα προηγούμενα κεφάλαια αναφέρθηκαν λεπτομερώς τα πλεονεκτήματα και μειονεκτήματα των διαφόρων στρατηγικών δειγματοληψίας. Μια όμως συγκριτική ανάλυση θα δώσει μια πιο χρήσιμη πληροφορία. Αυτό που συνήθως επιθυμούμε από έναν εκτιμητή είναι η όσο το δυνατό μεγαλύτερη ακρίβειά του κρατώντας φυσικά το κόστος της μελέτης σταθερό.

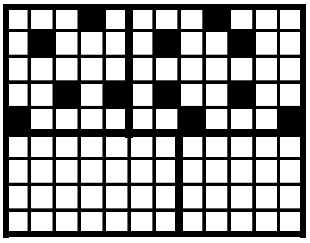
Όπως είπαμε στο πρώτο κεφάλαιο ο σχεδιασμός της δειγματοληψίας πρέπει να γίνεται με τρόπο που να αποφεύγονται τα "ακραία" δείγματα. Και ακραία είναι τα δείγματα των οποίων τα χαρακτηριστικά απέχουν πολύ απ'αυτά του πληθυσμού. Στο σχήμα 7.1 παρουσιάζεται η κατανομή στοιχείων δείγματος σε ένα πληθυσμό σύμφωνα με τις διάφορες στρατηγικές δειγματοληψίας. Για να είναι συγκρίσιμες οι μέθοδοι, το μέγεθος του δείγματος είναι σταθερό (1/9 του πληθυσμού). Από την πρώτη ματιά φαίνεται πως η στρωματοποιημένη και η συστηματική δειγματοληψία "καλύπτουν" καλύτερα τον πληθυσμό. Άρα οι δυο αυτές στρατηγικές μειώνουν το ρίσκο ακραίου δείγματος. Θα πρέπει λοιπόν να τις προτιμούμε. Όμως το αποτέλεσμα της δειγματοληψίας εξαρτάται και από τα βασικά χαρακτηριστικά του πληθυσμού και κυρίως τη διασπορά του. Όσο πιο ετερογενής είναι ο πληθυσμός τόσο πιο μεγάλο είναι το ρίσκο ακραίου δείγματος. Φυσικά την ίδια διασπορά μπορούμε να την έχουμε με διάφορους τρόπους. Μπορούμε δηλαδή να έχουμε δυο μεγάλες συναθροίσεις των στοιχείων του πληθυσμού ή πολυάριθμες μικρού μεγέθους. Η απόφαση λοιπόν για την επιλογή μιας στρατηγικής δεν είναι και τόσο απλή.



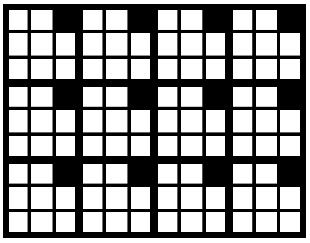
Απλή τυχαία δειγματοληψία



Στρωματοποιημένη δειγματοληψία



Δισταδιακή δειγματοληψία



Συστηματική δειγματοληψία

ΣΧΗΜΑ 7.1 *Παρουσίαση της κατανομής δειγματοληπτικών μονάδων σε ένα πληθυσμό σύμφωνα με τις διάφορες στρατηγικές (το μέγεθος του δείγματος παραμένει το ίδιο).*

Φυσικά εκτός από την ακρίβεια των εκτιμήσεων ο γενικότερος σκοπός της μελέτης καθώς και τα βασικά δομικά χαρακτηριστικά του πληθυσμού στόχου είναι καθοριστικά στην επιλογή της μεθόδου. Για παράδειγμα αν από τη φύση του ο πληθυσμός είναι χωρισμένος σε υποσύνολα (ομάδες) τότε η επιλογή μιας στρατηγικής που να εκμεταλλεύεται αυτό το γεγονός είναι σχεδόν αναπόφευκτη. Μπορεί επίσης η ίδια η μελέτη να επιζητά στοιχεία ή χαρακτηρισμό των υποσυνόλων ή ακόμα να επιθυμεί να δώσει περισσότερο βάρος σε κάποιες ομάδες.

Το δεύτερο στοιχείο που παίζει καθοριστικό ρόλο στην επιλογή της στρατηγικής είναι το κόστος δειγματοληψίας. Η έννοια του κόστους είναι σύνθετη. Εκτός από το κόστος ανάλυσης του επιλεγμένου στοιχείου, που είναι το ίδιο για όλες τις μεθόδους (φθάσαμε δηλαδή στο άτομο, τη βασική μονάδα, και μετράμε κάποιο χαρακτηριστικό του), έχουμε και το κόστος μετακίνησης, το κόστος δημιουργίας της λίστας των στοιχείων του πληθυσμού (όταν αυτό είναι αναγκαίο), το κόστος αναγνώρισης των ομάδων του πληθυσμού (στρώσεις, πρωτογενείς μονάδες, σειρά εμφάνισης των στοιχείων), και τέλος το κόστος αναγνώρισης και προσέγγισης των προς μέτρηση στοιχείων. Για παράδειγμα, στην περίπτωση μελέτης της αφθονίας ενός θαλάσσιου οργανισμού σε μια περιοχή συχνά χρησιμοποιούμε μια στρωματοποιημένη δειγματοληψία με βάση τοπογραφικά και/ή υδρολογικά χαρακτηριστικά. Το συνολικό κόστος της μελέτης αποτελείται από:

1. το κόστος ορισμού των στρώσεων. Για τον καθορισμό τους είναι αναγκαία στοιχεία που είτε υπάρχουν διαθέσιμα σε διάφορες υπηρεσίες και πρέπει να συγκεντρωθούν είτε πρέπει να συλλεχθούν στα πλαίσια της μελέτης
2. το κόστος μετακίνησης στα διάφορα σημεία της δειγματοληψίας (σταθμούς) που στη συγκεκριμένη περίπτωση απαιτεί πλωτό μέσο που συνήθως έχει και μεγάλο κόστος μετακίνησης και συντήρησης
3. το κόστος αναγνώρισης του προεπιλεγμένου για δειγματοληψία σημείου (γεωγραφικές ή άλλες συντεταγμένες που ανάλογα με την επιθυμητή ακρίβεια απαιτούν εξειδικευμένο ηλεκτρονικό υλικό συνήθως υψηλού κόστους)
4. το κόστος συλλογής του δείγματος (αλιευτικό εργαλείο ή εξειδικευμένη συσκευή για την συλλογή π.χ. βενθικών οργανισμών ή ιζήματος συγκεκριμένου πάχους)
5. το κόστος επεξεργασίας και συντήρησης του δείγματος (δοχεία απλά ή εξειδικευμένα, χημικά συντηρητικά, χώρος αποθήκευσης)
6. το κόστος ανάλυσης του δείγματος (αναγνώριση των επιθυμητών οργανισμών, μετρήσεις
7. το κόστος αρχειοθέτησης της πληροφορίας
8. και τέλος το κόστος ανάλυσης και παρουσίασης των δεδομένων.

Οι δαπάνες 4 έως 8 μπορούν να θεωρηθούν κοινές για όλες τις στρατηγικές και εξαρτώνται μόνο από το σκοπό της δειγματοληψίας. Έτσι η επιλογή της μεθόδου θα επηρεάσει τις δαπάνες 1-3.

Το τρίτο στοιχείο που παίζει συχνά καθοριστικό ρόλο για την επιλογή της στρατηγικής είναι κάποια πρακτικά προβλήματα που συχνά είναι ανυπέρβλητα και επιβάλλουν σχεδόν μια από τις στρατηγικές. Για παράδειγμα ο ακριβής κατάλογος όλων των ατόμων ενός φυσικού πληθυσμού είναι αδύνατος και συνεπώς τεχνικές που βασίζονται στην επιλογή τυχαίων ατόμων από το σύνολο του πληθυσμού δεν μπορούν να εφαρμοσθούν. Άλλα τέτοια προβλήματα είναι για παράδειγμα η ανικανότητα να αποφασισθεί με σιγουριά σε πια στρώση ανήκει ένα στοιχείο του πληθυσμού (λόγω ασάφειας στον ορισμό των στρώσεων ή περιορισμένης ακρίβειας πληροφορία για την προέλευση του στοιχείου).

Ας εξετάσουμε λοιπόν συγκριτικά τις διάφορες μεθόδους.

Στις επόμενες παραγράφους θα θεωρήσουμε σαν αναφορά την απλή τυχαία δειγματοληψία. Θα αναφερθούν επίσης τύποι και στοιχεία που πηγάζουν από τους βασικούς εκτιμητές που έχουν αναφερθεί στα προηγούμενα κεφάλαια αλλά που η απόδειξή τους ξεπερνά τα όρια αυτού του συγγράμματος. Η ανάπτυξη αυτών των θεμάτων μπορεί να βρεθεί σε εξειδικευμένα άρθρα και βιβλία με κύρια αναφορά αυτό του Cochran (1977).

Εάν ο πληθυσμός έχει χωρισθεί σε L στρώσεις και το δείγμα κατανεμηθεί στις στρώσεις σύμφωνα με αναλογική κατανομή (proportional allocation, $n_h/n=N_h/N$ ή $n_h=w_h n$) τότε η διασπορά της μέσης τιμής είναι

$$\mathrm{Var}(\bar{Y}_{\mathrm{strat-prop}}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{L} w_h S_h^2$$

(επιλέγουμε την περίπτωση της αναλογικής κατανομής γιατί οι τύποι απλουστεύονται).

Η διασπορά της μέσης τιμής της απλής τυχαίας δειγματοληψίας από τον ίδιο πληθυσμό και με το ίδιο συνολικό μέγεθος δείγματος ν μπορεί να γραφεί (για απλούστευση θεωρούμε τα $1/N_h$ αμελητέα, τύποι γενικής εφαρμογής βρίσκονται στον Cochran 1977, σελ. 99-101)

$$\mathrm{Var}(\bar{Y}_{\mathrm{rand}}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{L} w_h S_h^2 + \frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{L} w_h (\bar{Y}_h - \bar{Y})^2$$

Από τους δυο αυτούς τύπους φαίνεται ότι η διασπορά της μέσης τιμής της απλής τυχαίας δειγματοληψίας είναι μεγαλύτερη της διασποράς της στρωματοποιημένης κατά

$$\frac{1}{n}\left(1-\frac{n}{N}\right)\left(\sum_{h=1}^{L} w_h (\bar{Y}_h - \bar{Y})^2\right)$$

αυτή η ποσότητα είναι ≥0. Η ποσότητα αυτή είναι ίση με το 0 μόνο όταν οι μέσες τιμές όλων των στρώσεων είναι ίσες με τη μέση τιμή του πληθυσμού (πράγμα σπάνιο στην πράξη). *Από τους τύπους αυτούς φαίνεται καθαρά ότι όσο πιο ομοιογενείς είναι οι στρώσεις στο εσωτερικό τους και όσο πιο πολύ διαφέρουν μεταξύ τους (άρα και οι* $\overline{Y}_i$ *θα διαφέρουν πολύ από τη μέση τιμή* $\overline{Y}$*) τόσο μεγαλύτερο είναι το κέρδος της στρωματοποίησης*. Κατά συνέπεια η διασπορά της στρωματοποιημένης δειγματοληψίας είναι κατά μέσο όρο μικρότερη από αυτή της απλής τυχαίας. $\mathrm{Var}(\overline{Y}_{\text{strat-prop}}) \le \mathrm{Var}(\overline{Y}_{\text{rand}})$ όταν οι ποσότητες $1/N_h$ μπορούν να θεωρηθούν αμελητέες (κάτι που ισχύει συχνά στη μελέτη των φυσικών πληθυσμών)..

Η σύγκριση ανάμεσα στην στρωματοποιημένη και την πολυσταδιακή δειγματοληψία δεν είναι τόσο εύκολη. Για να έχουμε μια όσο το δυνατό συμβατότερη εικόνα θα θεωρήσουμε την περίπτωση της δισταδιακής δειγματοληψίας (οι στρώσεις αντιστοιχούν στις πρωτογενείς μονάδες). Μια από τις βασικές δυσκολίες είναι και τρόπος με τον οποίο μοιράζεται η δειγματοληπτική προσπάθεια στα διάφορα επίπεδα. Το ίδιο μέγεθος τελικού δείγματος μπορεί να επιτευχθεί παίρνοντας λίγες πρωτογενείς μονάδες και πολλές δευτερογενείς στο εσωτερικό τους ή το αντίθετο. Έτσι ένας πρακτικός κανόνας επιλογής είναι: *όσο μικρότερες είναι οι διαφορές ανάμεσα στις μέσες τιμές (*$\bar{y}_h$*) και τις διασπορές (*s_h^2*) των υποσυνόλων του πληθυσμού (το κάθε υποσύνολο δηλαδή αποτελεί ένα αντίγραφο μινιατούρα του πληθυσμού) τόσο πιο αποδοτική είναι η δισταδιακή δειγματοληψία, αντίθετα, όπως ήδη αναφέρθηκε, όσο μειώνεται η διασπορά στο εσωτερικό των υποσυνόλων (intra-group variance) και αυξάνονται οι διαφορές μεταξύ τους (inter-group variance) τόσο ενδείκνυται η στρωματοποίηση*.

Η συστηματική δειγματοληψία μπορεί να καλύπτει καλύτερα το δείγμα όπως φαίνεται και από το σχήμα 7.1 *όμως είναι περισσότερο ακριβής από την απλή τυχαία δειγματοληψία μόνο όταν η μέση διασπορά στο εσωτερικό των συστηματικών δειγμάτων είναι μεγαλύτερη από τη διασπορά του πληθυσμού,* δηλαδή όταν (Cochran, 1977, σελ. 208)

$$S^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2}{N-1} < \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2$$

Κατά συνέπεια η συστηματική δειγματοληψία είναι πιο ακριβής από την απλή τυχαία όταν τα συστηματικά δείγματα είναι ιδιαίτερα ετερογενή. Αυτό είναι αυτονόητο γιατί αν υπάρχει μικρότερη ετερογένεια στο εσωτερικό του συστηματικού δείγματος σε σχέση με αυτήν του πληθυσμού τότε τα στοιχεία του δείγματος δεν κάνουν τίποτε άλλο από το να επαναλαμβάνουν λίγο-πολύ την ίδια πληροφορία και κατά συνέπεια να βρίσκονται μακριά από την

πραγματικότητα. Φυσικά όπως αναφέρθηκε στο αντίστοιχο κεφάλαιο όταν η διασπορά των συστηματικών δειγμάτων είναι ιδιαίτερα μικρή η συστηματική συλλογή πρέπει να αποφεύγεται ή να λαμβάνονται σοβαρές προφυλάξεις πριν την εφαρμογή της.

Από όλα αυτά φαίνεται ότι δεν υπάρχει καθαρή ιεράρχηση των μεθόδων όσον αφορά στη διασπορά των εκτιμητών τους, αλλά η ακρίβειά τους εξαρτάται από τα χαρακτηριστικά του πληθυσμού. Οποιαδήποτε λοιπόν πληροφορία που σχετίζεται με τη δομή του πληθυσμού από παλαιότερες μελέτες, από συγγενείς πληθυσμούς, από τη βιβλιογραφία κ.λ.π. πρέπει να χρησιμοποιείται διότι θα επιτρέψει την επιλογή της σωστής στρατηγικής και συνεπώς θα μειώσει την διασπορά του εκτιμητή.

Από την πλευρά του κόστους τα πράγματα είναι λίγο πιο ξεκάθαρα. Η συστηματική δειγματοληψία έχει το μεγαλύτερο κόστος μετακίνησης (όταν πρόκειται για πληθυσμούς διεσπαρμένους στο χώρο) ακολουθούμενη από την στρωματοποιημένη, την απλή τυχαία και τέλος την πολυσταδιακή. Ένα επιπλέον κόστος προέρχεται από την δημιουργία του καταλόγου των στοιχείων του πληθυσμού με σκοπό την τυχαία επιλογή κάποιων απ'αυτά. Μόνο η πολυσταδιακή δειγματοληψία περιορίζει αυτό το κόστος με τη δημιουργία καταλόγου των στοιχείων μόνο των πρωτογενών μονάδων που έχουν επιλεγεί. Σ'αυτόν τον τομέα η συστηματική δειγματοληψία απαιτεί την απόδοση ενός αύξοντος αριθμού στα στοιχεία του πληθυσμού ώστε να επιλέγεται ένα κάθε p (το βήμα της δειγματοληψίας). Στην περίπτωση των φυσικών πληθυσμών αυτό συχνά είναι αδύνατο κι έτσι περιοριζόμαστε στην σειρά εμφάνισης των ατόμων και παίρνουμε ένα κάθε p. Φυσικά η σειρά αυτή πρέπει να είναι ξεκάθαρη, π.χ. ένα σμήνος πουλιών που εμφανίζεται ξαφνικά ποια είναι η σειρά εμφάνισης των ατόμων;. Σ'αυτή όμως την περίπτωση τελικά δεν μπορούμε να γνωρίζουμε από πριν το συνολικό μέγεθος του δείγματος.

Τέλος οι διάφορες μέθοδοι εκμεταλλεύονται με συγκεκριμένο τρόπο και σε διαφορετικό βαθμό την διαθέσιμη πληροφορία από άλλες παρόμοιες μελέτες από παλαιότερες παρατηρήσεις ή άλλη συσσωρευμένη γνώση. Για παράδειγμα η απλή τυχαία δειγματοληψία καθώς και η συστηματική δεν χρησιμοποιούν κανένα στοιχείο σχετικό με τον πληθυσμό στόχο. Πρέπει λοιπόν να γίνει κατανοητό ότι η χρησιμοποίηση πληροφορίας για το σχεδιασμό μιας δειγματοληπτικής στρατηγικής δεν είναι τίποτε άλλο παρά χρήμα που επενδύεται στη μελέτη (διότι αυτή η πληροφορία όταν είχε αποκτηθεί είχε απαιτήσει κάποια δαπάνη) και που έχει σαν αποτέλεσμα την αύξηση της ακρίβειας της εκτίμησης.